An Improvised Sub-Document Based Framework for Efficient Document Clustering

Muhammad Qasim Memon[1,2], Jingsha He[1,3], Yu Lu[2], Nafei Zhu[1], Aasma Memon[4]

[1] Faculty of Information Technology, Beijing University of Technology, China
[2] Advanced Innovation Center for Future Education, Faculty of Education, Beijing Normal University, China
[3] College of Computer and Information Science, China Three Gorges University, China
[4] School of Economics and Management, Beijing University of Technology, China
memon_kasim@bnu.edu.cn, jhe@bjut.edu.cn, luyu@bnu.edu.cn, znf@bjut.edu.cn, kaasma.bjut@gmail.com

Abstract

Document clustering, which is used for topic discovery and similarity computation, has received a great deal of attention in text data management. Methods that have been adopted in traditional clustering, particularly for multi-topic documents, are not viable because the contents that are distinguished by the sub topical structure may not be pertinent across the entire documents. In this paper, a sub-document based framework for clustering multiple documents is proposed in which LDA is used for document segmentation. The proposed improvised framework is a two-way approach to address the clustering problem. First, instead of applying a clustering algorithm to the entire data sets, documents are partitioned into cohesive sub-documents along topic boundaries through text segmentation to establish a two-level representation of text data, i.e., topics and words. Second, the proposed framework is compared to existing clustering methods, both traditional and segment based clustering through different clustering algorithms using the F-measure as the measurement metric. In addition, various real-time data sets that contain multi-topic documents are applied to validating the clustering algorithms through the proposed sub-document based framework. Each sub-document is clustered within a document and the resulting clusters are further clustered across the documents. Experimental results show that the proposed framework outperforms existing clustering approaches in terms of the F-measure as well as efficiency at least 73% with LDA segmentation and bisecting LDA in comparison to TextTiling.

Keywords: Clustering algorithms, Text analysis, Text mining, Information retrieval, Data mining

1 Introduction

Text clustering is the process of classifying a collection of text units in a meaningful, congregated manner to find relevant topics with high intra class similarity. These text units (known as documents) are assembled based on topic similarity or categorically defined to explore informative data with comparative ease. The literature in this regard suggests that the two clustering approaches of agglomerative hierarchical clustering (AHC) and partitional clustering are already well established and successful in the domain of document clustering. In AHC, each document is classified according to topic similarity into clusters and computed using distance functions [1]. Partition clustering provides overlapping solutions of clusters using k-means algorithms and other probabilistic algorithms. However, the major concern is how to develop a document clustering approach that would evaluate each document that is explicitly related to different topics. Document clustering is interconnected with multi-topics models, where the best topic model is representative of viable attributes, such as a decrease in the similarity measure and the sufficient recognition of the collected structure of the corpus. The best topic model should be coherent and semantically strong with the most essential part of the model to find the most topics from the corpus. Existing methods have been found to be weak on the analysis of multi-topics documents, portraying subpar results with little or no connection to the topics' similarity to determine the perspective domains [2]. Traditional clustering of multi-topic documents involves different approaches that may produce overlapping clustering solutions, such as fuzzy clustering, clustering based on generative models and ensemble subspace clustering [3-5]. Experimental results obtained by applying our proposed framework on two multi-topic datasets (in agreement with different algorithms) have been compared to those obtained using existing methods, such as multi-document segment based clustering [6] and multi-document segment based clustering using TextTiling [7].

The objective of this research work is to resolve the limitations observed in the existing clustering

[*] Corresponding Author: Muhammad Qasim Memon; E-mail: memon_kasim@bnu.edu.cn

DOI: 10.3966/160792642019072004018

approaches by applying the proposed clustering framework in which documents are partitioned into sub-document sets (i.e., clusters comprised of individual topics within each cluster as text units). Our work in this paper has the following main contributions in comparison to existing approaches:

- The proposed framework emphasizes on topic modeling to improve segmentation using LDA manipulated with clustering algorithms.
- The proposed framework would extract sub-documents using LDA based segmentation in which sub-documents are identified by computing the rate of error in segmentation (in terms of P_k) based on topics and words.
- Sub-documents are labelled with topic information in the proposed framework using LDA segmentation.
- The proposed framework is first evaluated on two standard data sets. Later, different real-time data set containing multi-topic documents are designed to perform an in-depth demonstration that could improve the performance of bisecting Sk-Mean and bisecting LDA.

The remainder of this paper is organized as follows. In Section 2 briefly explains related work on document clustering for multi-topic documents. Section 3 provides the background of document representation models in our framework. Section 4 contributes to the interior declaration of clustering in sub-document based clustering. Section 5 presents experiment results and their analysis with respect to our proposed framework. Finally, Section 6 concludes the paper.

2 Related Work

A number of clustering algorithms have been proposed, including partition clustering (which is fuzzy clustering aimed to provide overlapping solutions) [7] and probabilistic generative models [8]. Probabilistic models define each document as a mixture of topics distributed across the terms [9]. Probabilistic Latent Semantic Analysis (PLSA) is an extension of Latent Semantic Analysis (LSA) in which each document is represented as a mixture of topics based upon a distribution through gathered terms [10]. Latent Dirichlet allocation (LDA) is also considered as a mixed model consisting of entire documents, corpuses and terms [11].

Text segmentation is used in information retrieval application to divide the given text data into multiple cohesive parts (called segments) with various topics and subtopics [12]. It can enhance the performance of the user's retrieval experience by prevailing associated parts of text. Different methods [4, 8] have been used to find the thematic parts of documents, which are identified through boundaries (known as segments) and lexical chain information based upon the associated words in cohesive segments. Thus, in each segment, boundaries are subject to changes in vocabulary [13]. However, these methods generate the segment boundaries with the disadvantage of the words being repeated throughout the process of segmenting the text data. To overcome this issue, different approaches are proposed [14-15]. An inclusion of ontology-based document clustering could be useful to exploit the semantic relation between words in order to improve the clustering quality. An ontology-based general weighting schema framework proposed by [16], which incorporates different map functions on to a based taxonomy. This taxonomy can be any sort of dataset or database such as WordNet. Similarly, an e-Learning domain specific ontology-based document clustering is compared to the traditional clustering by defining weighting scheme comprised of co-occurrences of words and with weights of relations among terms [17]. However, there still exist different issues, such as retrieving word semantics from texts, synonym and polysemy, appropriate declaration of clusters and high dimensionality. In order to remove these issues the integration of WordNet and lexical chain were attempted to generate clusters with accurate assessment of terms for word sense disambiguation [18]. Existing document clustering methods [19-20] such as ontology-based [21] in combination with LDA [22], LSA [23] and fuzzy clustering [24], ensemble sup-space clustering [3], topic identification based clustering [25], agglomerative hierarchical clustering [1, 9] and probabilistic model based clustering [8] were proposed in order to improve the clustering quality in terms of accuracy and efficiency. The above-mentioned clustering methods were mostly biased to cluster each document as a single text unit and found less effective to provide efficient and accurate clusters. Whereas, topic modeling and document segmentation methods that coincide document segmentation and document clustering could be converged based on our proposed sub-document based framework in order to improve the quality of clusters rendering accurate and efficient clustering results.

3 Document Representation and Notations

3.1 Sub-document Clustering

Consider a set D of documents $d \in D$ that consist of intransitive sub-documents *sd* and they are partial to each other. For an *Sd* set of sub-documents from document d, the sub-document set from all documents is denoted as $\boldsymbol{S} = \cup_{d \in D} Sd$. Table 1 presents all the notations that are used in this paper.

3.2 Sub-document Set Representation

The clustering of sub-documents using k-way clustering is performed by the spherical k-means algorithm (Sk-Means) to produce a disjointed clustering

Table 1. Notations

Symbol	Description
D	Collection of documents
d	Document
sd	Sub-document
Sd, Sd_d	Sub-document set, sub-document in a document
S, S_d	Collections of sub-document sets, set of sub-document sets in a document
N_D	Number of documents in D
N_{Sd}	Number of sub-documents in Sd
N_S	Number of sub-document sets in S
$N_{\mathbf{D}}$	Number of documents in D
$\mathcal{C}_S$	Sub-document set clustering
$\mathcal{C}$	Document clustering
C	Document cluster
h	Topic label in D
h_d	Topic label in *d*
k	Average count of sub-document sets
sdtf-idf	Sub-Document Set Term Frequency-Inverse Document Frequency
sdtf-isdf	Sub-Document Set Term Frequency-Inverse Sub-document Frequency
sdtf-isdsf	Sub-Document Set Term Frequency-Inverse Sub-document Set Frequency
α	Parameters of topic Dirichlet prior
β	Parameters of word Dirichlet prior
W	Set of words w_i assigned to W
z	z_i latent topics assign to words in W
ϵ	Probability of topic $z = k \in d$

solution based on partitional clustering [7]. The TF-IDF is applied to the document, sub-document and sub-document set using the compensating term weighting functions similar to the tf-idf. We assume tf(w,*Sd*) where w is an index term and *Sd*∈*S* is a sub-document set. The frequency functions of the document, sub-document and sub-document-set are respectively computed by the sdtf-idf, sdtf-isdf and sdtf-isdsf as follows:

$$stdf - idf(w, Sd) = tf(w, Sd) \times \log(\frac{N_D}{N_D(w)}) \quad \mathbf{(1)}$$

$$sdtf - isdf(w, Sd) = tf(w, Sd) \times esp(\frac{N_{Sd(w)}}{N_{Sd}}) \times \log(\frac{n_S}{n_S(w)}) \quad \mathbf{(2)}$$

$$sdtf - isdf(w, Sd) = tf(w, Sd) \times \log(\frac{N_S}{N_S(w)}) \quad \mathbf{(3)}$$

where N_D is the number of documents in D and $N_D(w)$ is portion of D that includes w. N_{Sd} contains sub-documents of Sd and N_s contains sub-documents of S. $N_{Sd}(w)$ and $n_s(w)$ are the respective distributions of Sd and S that include w. N_s is the number of sub-documents sets in S and $N_s(w)$ is the distribution of S including w. The exponential factor is introduced for the sub-documents to enhance the frequency of terms within the sub-document set.

3.3 Document Clustering Representation

Each cluster of the sub-document set is replaced in its respective original document by performing disjointed clustering C_s of the sub-document sets to form overlapping clustering solutions. A sub-document set is supposed to have one particular topic, while a document can be designated to have multi-topics in the entire collection of documents, such as C_s = $\{C_1^{(d)}, ..., C_h^{(d)}\}$ of clusters over **D**. The pseudo code of sub-document-based clustering using the LDA text segmentation algorithm is given below:

Algorithm. Sub-document based clustering using LDA segmentation

```
1. TS : Text Segmentation algorithm
   i. LDA method
   ii. TextTiling method
2. LDA method: for each possible sub-document, $sd_i$
   Find Θ, with
      $\theta_{sd_i t} \leftarrow \frac{1}{n_i} \sum_{v=1}^{V} \frac{C_{wIv} \theta_{sd_i t'} \phi_{tv}}{\sum_{t'=1}^{T} \theta_{sd_i t'} \phi_{t'v}}$
   b. Find its likelihood, with
      $P(W_i \mid \Theta, \Phi)\theta_{sd_i t} = \prod_{v=1}^{v} [\sum_{t=1}^{T} (\theta_{sd_i t} \phi_{tv})]^{Cw_i v}$
   c. Sub-document likelihood is computed by its rating:
      $\log P(P \mid Sd_i) = \log(W_i \mid \Theta, \Phi)$
3. SC: soft partition clustering
4. DC: hard partition clustering
5. Sub-document, $S \leftarrow \varnothing$
6. For all $d \in D$, do
7.   $S_d \xleftarrow{TS}$, retrieve sub-documents
8.   $Sd \leftarrow Sd \cup S_d$
9.   $S \leftarrow \varnothing$
10. For all retrieved, $S_d \in Sd$, do
11.   $S_d \xleftarrow{SC} (S_d)$
12.   $S \leftarrow S \cup S_d$
13.   $\mathcal{C}_S \xleftarrow{DC} (S)$        // cluster sub-documents sets
```

4 Methodology

In the proposed framework, sub-document based clustering is performed on two datasets known as

RCV1 and 20 Newsgroups [26-27]. In experiment 1, we consider 21850 documents from the RCV1 dataset after filtering the short structured news related to finance and politics. Therefore, a document is assumed to have at least one paragraph of three lines with few sentences while the topic sentences double the number of associated topics. 20 Newsgroups dataset is a collection of approximately 18,828 documents that are evenly partitioned across 20 different newsgroups. Each newsgroup corresponds to different topics, such as computers, politics, and more. The datasets are restricted with a couple of limitations, which are referred to as pre-requisite for clustering. Therefore, document consists of no more than three topics and must possess two percentages of topics. Data preprocessing is applied on dataset's stop-words, removing strings of digits and words stemming to mature data analysis. Datasets used in experiment 1 are briefly explained in Table 2. As previously stated, the clustering of sub-documents in each document is performed by the Sk-Means, LDA and OSk-Means. However, the maximum number of iterations for clusters in our case is restricted to of 60. The exploitation of LDA and TextTiling based segmentation to identify the sub-documents in a document is shown in Figure 1.

Table 2. Datasets used in the experiment

Dataset	#docs	#topic labels	#terms	#docs per topic	#sub-docs per doc
RCV1	6456	23	37991	280.7	5.9
20 Newsgroups	5644	20	25553	282.2	5.4

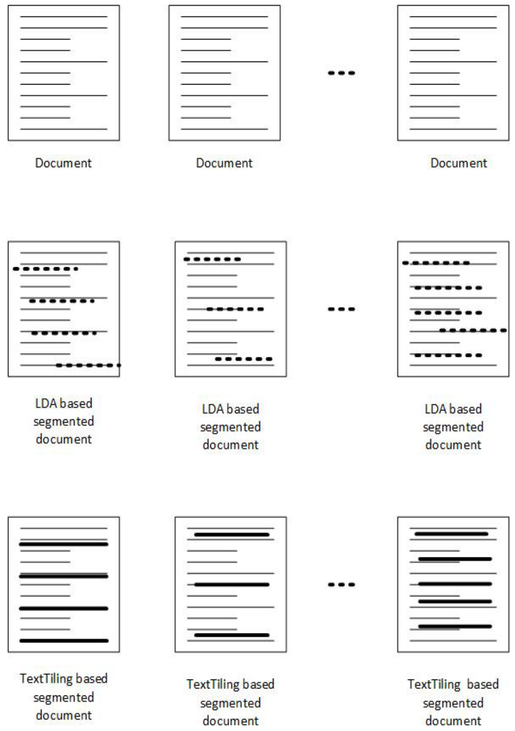

Document
Document
Document
LDA based segmented document
LDA based segmented document
LDA based segmented document
TextTiling based segmented document
TextTiling based segmented document
TextTiling based segmented document

Figure 1. Representation of sub-documents using LDA and TextTiling

4.1 Evaluation Methods

4.1.1 Document Segmentation

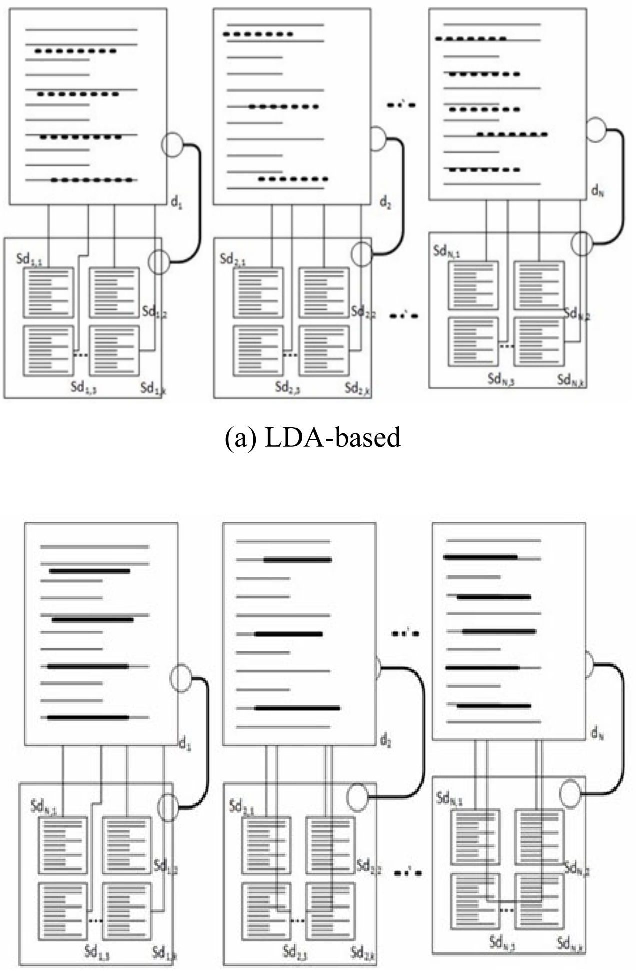
In TextTiling algorithm, sub-documents are identified within documents by setting parameters in [7]. The recommended values for the token size and text unit were 20 and 6 ÷ 10, respectively. However, we varied the default values for both the size of the text unit. Token sequence size between 2 to 12 and approximately ± 14, respectively. In LDA, each dataset was divided into subsequent subsets based upon sentences in the sub-document. First three subsets ("S1," "S2", "S3") were divided in a sub-document that depended on the number of sentences in a sub-document. The fourth subset ("S1-S3") contained whole dataset and was named "EntireText". For each subsequent subset (Excluding fourth), each document had 20, 50 and 100 sub-documents to emulate the wider spectrum of text segmentation. Vocabulary size for the training set of RCV1 and 20 Newsgroups was approximately 97K and 82K, respectively. The number of word tokens for both datasets were approximately 3.6 M and 2.0 M, respectively. The Dirichlet prior value of α = 0.01 was taken based on its positive impact. The number of topics were taken as T=40 to balance the computational complexity of the segmentation algorithm. Computational costs were enhanced as T increased while training the LDA model. The results of comparing LDA with TextTiling segmentation algorithm [7] are reported in terms of precision, recall and F-measure. These measures are declared as:

$$\text{Precision} = \frac{\text{no of estimated boundaries that are actual}}{\text{no. of estimated boundaries}} \tag{4}$$

$$\text{Recall} = \frac{\text{no of estimated boundaries that are actual}}{\text{no. of actual boundaries}} \tag{5}$$

$$\text{F}^{\text{M}} = \frac{2 * \text{precision} * \text{recall}}{\text{precisionu} + \text{recall}} \tag{6}$$

Figure 2, shows the no cross clustering through sub-document using LDA and TextTiling, respectively. In LDA, sub-documents associated to each document are segmented at different places and appeared in diverse manner compared to sub-documents using TextTiling. Since, Sub-documents are in constant and fixed position (in contiguous blocks) if compared to original document using TextTiling. On the contrary, sub-documents are identified based on topics and words and their segmentation boundary is estimated by computing segmentation error using LDA.



(a) LDA-based

(b) TextTiling-based

Figure 2. No cross sub-document clustering using sub-document and sub-document set

4.1.2 Cross Clustering Model Selection

Mapping of sub-document set cross clustering (i.e. clusters) to a respective document (disjoint clustering) is shown in Figure 3. Sub-document set cross clustering resulting in clusters associating exactly one topic across different documents in shape of sub-documents. Resultant clusters are further clustered in overlapping fashion in order to associate topic similarity across different documents. For example, $\{C_1, C_2, \dots, C_k, C_N\}$ represents clusters that contain sub-document from different documents using LDA segmentation.

We stick with most common criterion F-measure for clustering the quality of the documents in terms of precision and recall. Take a set D of documents $C = \{C_1, \dots, C_h\}$ and $c = \{c_1, \dots, c_k\}$, where C is the classification of document **D** and C is clustering across **D**. Consider a pair P $(C_j,\ c_i)$ and (P_{ij}) that represent the division of c_j such that $P_{ij} = \frac{|C_j \cap c_i|}{|C_j|}$. Meanwhile,

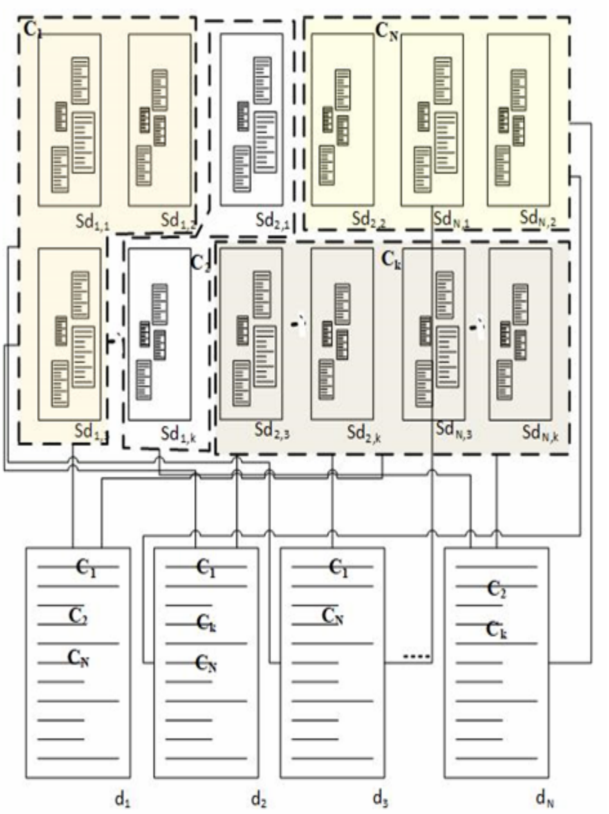


Figure 3. Sub-document sets mapping into a document

consider a pair (c_j, c_i) and (R_{ij}) represents the division of c_i such that $R_{ij} = \frac{|C_j \cap c_i|}{|C_i|}$. To measure the quality of C with C, F-measure computed by using the harmonic means between precision and recall. The F-measure is introduced by considering macro average F^M and micro average F^μ. Macro average assigns equal weight to each class and micro average gives equal weight to each sub-document of a document. Micro F-measure is used to estimate quality of clustering and mainly converges during the analysis of experimental results. Thus,

$$F^M = 2PR(P+R) \tag{7}$$

and

$$R = \frac{1}{h}\sum_{i=1}^{h} \max_{j=1\ldots k\{R_{ij}\}} \tag{8}$$

$$F^\mu = \sum_{i=1}^{h} (\frac{|c_i|}{|\boldsymbol{D}|}) \max_{j=1\ldots k\{F_{ij}\}} \tag{9}$$

wher $F_{ij} = 2P_{ij} \times R_{ij}(P_{ij} + R_{ij})$. Since, clustering algorithm depends on random initialization and number of iterations. Results correlate upon these multiple computed runs. The specified parameter values of overlapping clustering solutions may vary on the algorithm. For LDA, probability threshold value ranges from 5.00E - 05 to 0.30. However, each dataset appears to have various ranges of scores in the F-measure i.e.

normalized by the maximum F-measure. Overlapping solutions of remaining datasets are computed by parameter values over the maximum averaged micro F-measures across the dataset.

5 Results and Analysis

5.1 Experiment 1: Results and Discussions

5.1.1 Document Segmentation

The document segmentation conducted using LDA segmentation in this research and was compared with TextTiling algorithm. Table 3 shows the performances of LDA in terms of P_k , which is likelihood that arbitrarily chosen sentences are considered as kth sentences. The table data suggests that the highest value of P_k indicates more accurate text segmentation. Performances of TextTiling and LDA methods for RCV1 and 20 Newsgroups in terms of the F-measure and P_k are shown in Table 4 and 5, respectively. The higher value of P_k and F-measure is taken as an indication of better performance.

Table 3. Performance of LDA in terms of P_k

Data set	S1	S2	S3	EntireText
RCV1	19.8	17.3	14.6	20.3
20 Newsgroups	17.6	15.5	12.1	16.4

Table 4 and Table 5 results shows computed values were represented as the average maximum micro F-measure. We selected half of the documents from each dataset and remaining documents were used for training set to find optimal parameters settings. TextTiling has exploited a sufficiently large number of sub-documents associated with original documents. However, it is less effective than the LDA based segmentation. The average F-measure score corresponded to standard deviations for both cross clustering and no cross clustering. While performing no cross clustering, the macro and micro F-measures were not better than cross clustering, resulting in major changes in F-scores. Total performance variation was found up to 1.1% with TextTiling and beyond 2% in the case of the LDA.

5.1.2 Sub-document Cross Clustering

Clustering of sub-documents was performed within each document using standard Sk-Means, LDA and OSk Means. Table 6 illustrates the characteristics of clustering achieved by the sub-document based clustering framework. Micro and macro F-measures could justify the efficacy of our approach applied in this research. Results suggest that sub-document across clustering performed better of approximately 18% on average than the other schemes in both clustering solutions. However, overlapping clustering algorithms outperformed non-overlapping with an average improvement of 6.3% and 7.8%. When sub-documents were clustered using h2-way, quality of clustering was enhanced as the number of sub-documents increased. Moreover, clustering solutions that were evaluated using sub-document clustering via the LDA and OSk-Means resembled overlapping clustering results were better than non-overlapping solutions (Sk-Means) and obtained better performances across the two datasets with 15.7% micro F-measure and a 16.7% macro F-measure using LDA.

Table 4. Performance of the TextTiling method

Data set	No cross clustering		Cross clustering		Total	
	F^{μ}	F^{M}	F^{μ}	F^{M}	F^{μ}	F^{M}
RCV1	.008	.009	.017	.016	.011	.011
20 Newsgroups	.006	.007	.015	.014	.009	.009

Table 5. Performance of the LDA method

Data set	No cross clustering		Cross clustering		Total	
	F^{μ}	F^{M}	F^{μ}	F^{M}	p_k	F^{M}
RCV1	.018	.021	.045	.042	20.3	0.43
20 Newsgroups	.018	.015	.026	.036	16.4	0.47

Table 6. Characteristics of sub-document clustering

Data set	Clustering Algorithm	h-way		h^2-way	
		F^{μ}	F^{M}	F^{μ}	F^{M}
RCV1	Sk-Means	.511	.501	.565	.542
	LDA	.559	.542	.587	.584
	OSk-Means	.531	.514	.582	.564
20 Newsgroups	Sk-Means	.506	.496	.545	.532
	LDA	.547	.531	.575	.571
	OSk-Means	.527	.505	.570	.567

5.1.3 Sub-document Set Clustering in Comparison to Traditional Clustering

Table 7 and Table 8 show the results using various sub-document based methods in comparison between LDA and TextTiling. These methods were performed through the h-way and h2-way using the Sk-Means, LDA and OSk-Means. Each clustering method was implemented using either cross or no cross level clustering. LDA was performed on sub-documents and sub-document sets to produce the h2-way clustering solution. Overall performance was observed to be worse in terms of the F-measure by Sk-Mean due to the generation of non-overlapping clustering solutions. Hence, this proves that it is not an effective algorithm to cluster multi-topic documents. Moreover, results could clearly suggest that the introduction of the LDA based segmentation outperformed the application of TextTiling. In addition, sub-document based schemes utilizing sub-document cross clustering produced better results than those obtained using proposed clustering methods [6]. The highest macro F-measure of 0.791 with an average improvement of 10.2% was observed on RCV1 dataset. This is compared to an average improvement of 11.2% in the case of the 20 Newsgroups dataset, which contained much smaller sub-documents within a document, which implies that sub-document of associated or related multiple topics across documents yielded better improvement. In terms of precision, our proposed clustering framework performed better than traditional document clustering methods with an average improvement of over 54%. The results obtained using our proposed framework also showed that the LDA and OSk-Means using the LDA based segmentation explicitly outperformed in comparison to the methods proposed elsewhere [6-7].

Table 7. Performance of sub-document based clustering of six clustering methods using TextTiling

Clustering Method	Clustering Algorithm	Data set							
		RCV1				20 Newsgroups			
		P	R	F^M	F^μ	P	R	F^M	F^μ
No Cross Document Clustering	Sk-Means	0.711	0.39	0.504	.414	0.679	0.35	0.462	.389
	LDA	0.489	0.67	0.565	.571	0.448	0.638	0.526	.541
	OSk-Means	0.461	0.641	0.536	.510	0.432	0.604	0.504	.478
No Cross Sub-document Clustering	Sk-Means	0.65	0.51	0.572	.545	0.611	0.479	0.537	.511
	LDA	0.634	0.558	0.594	.496	0.596	0.479	0.531	.444
	OSk-Means	0.647	0.544	0.591	.488	0.613	0.534	0.571	.426
Sub-document Cross Clustering	Sk-Means	0.61	0.528	0.566	.581	0.576	0.56	0.568	.539
	LDA	0.52	0.649	0.577	.529	0.486	0.623	0.546	.489
	OSk-Means	0.601	0.617	0.609	.501	0.578	0.582	0.58	.477
Document Cross Clustering	Sk-Means	0.81	0.371	0.509	.369	0.766	0.334	0.465	.340
	LDA	0.589	0.627	0.607	.506	0.547	0.57	0.558	.485
	OSk-Means	0.468	0.793	0.589	.623	0.43	0.745	0.545	.512
Sub-document set Cross Clustering	Sk-Means	0.702	0.431	0.534	.473	0.658	0.452	0.536	.437
	LDA	0.679	0.591	0.632	.467	0.64	0.56	0.597	.423
	OSk-Means	0.655	0.456	0.538	.431	0.632	0.419	0.504	.398
Sub-document & Sub-document set Cross Clustering	Sk-Means	0.654	0.679	0.666	.522	0.626	0.647	0.636	.481
	LDA	0.602	0.795	0.685	.601	0.576	0.691	0.628	.492
	OSk-Means	0.589	0.681	0.632	.520	0.563	0.652	0.604	.511

Table 8. Performance of sub-document based clustering of six clustering methods using LDA segmentation

Clustering Method	Clustering Algorithm	Data set							
		RCV1				20 Newsgroups			
		P	R	F^M	F^μ	P	R	F^M	F^μ
No Cross Document Clustering	Sk-Means	0.699	0.45	0.548	.430	0.658	0.4	0.498	.398
	LDA	0.502	0.665	0.572	.588	0.465	0.628	0.534	.546
	OSk-Means	0.477	0.621	0.54	.531	0.427	0.622	0.506	.504
No Cross Sub-document Clustering	Sk-Means	0.62	0.58	0.599	.564	0.623	0.484	0.545	.547
	LDA	0.666	0.561	0.609	.490	0.64	0.477	0.547	.459
	OSk-Means	0.688	0.57	0.623	.519	0.68	0.51	0.583	.478
Sub-document Cross Clustering	Sk-Means	0.579	0.565	0.572	.594	0.551	0.597	0.573	.560
	LDA	0.641	0.627	0.634	.533	0.605	0.586	0.595	.511
	OSk-Means	0.691	0.601	0.643	.525	0.647	0.571	0.607	.492
Document Cross Clustering	Sk-Means	0.74	0.394	0.514	.391	0.711	0.366	0.483	.357
	LDA	0.618	0.601	0.609	.526	0.58	0.57	0.575	.499
	OSk-Means	0.569	0.779	0.658	.654	0.555	0.714	0.625	.631
Sub-document set Cross Clustering	Sk-Means	0.639	0.526	0.577	.488	0.62	0.498	0.552	.451
	LDA	0.781	0.571	0.66	.499	0.739	0.549	0.63	.467
	OSk-Means	0.746	0.433	0.548	.509	0.709	0.406	0.516	.485
Sub-document & Sub-document set Cross Clustering	Sk-Means	0.691	0.688	0.689	.501	0.618	0.659	0.638	.476
	LDA	0.661	0.812	0.729	.518	0.654	0.727	0.689	.496
	OSk-Means	0.639	0.657	0.648	.621	0.603	0.638	0.62	.594

5.1.4 Performance Evaluation of Sub-document Based Clustering

Proposed sub-document based clustering performance was evaluated by measuring time costs and memory costs consumed by clustering algorithms. We conducted a set of experiments on five different datasets acquired from RCV1 and 20 Newsgroups collections by evolving random sampling with replacement method to evaluate the scalability on both small and large datasets using the five datasets of DS (1-5). The overview of five document datasets is presented in Table 9. Figure 4 shows the time costs of sub-documents using no cross clustering of five datasets DS (1-5) using TextTiling (on top) and the LDA segmentation methods (on bottom). These results showed that LDA clustering algorithm performed worse than the other two algorithms due to the independent clustering performed on each document. However, results obtained using LDA based segmentation achieved better performance with respect to time costs with an average improvement of 16% over TextTiling due to the huge number of sub-documents involved in the documents. Figure 5 shows time costs required for cross clustering among sub-document sets for five different datasets DS (1-5), which corresponded to the numbers of clustering solutions, such as 20, 40, 120 and 150-way using sub-document set representation. The LDA segmentation achieved better performance in terms of time costs and memory costs than that consumed by TextTiling algorithm because the topic distribution was not related to each sub-document in the sub-document set along with the sub-document boundary. Figure 6 shows three sets of results, which were obtained by performing h-way clustering for DS (1-5). In addition, results showed that performances for all sample datasets linearly decreased because each sample of the dataset was twice as large as the previous one. LDA clustering algorithm outperformed the other two schemes with an average improvement of over 24%. This meant that the sub document based method was capable of selectively recognizing the documents to properly cluster them compared to traditional document clustering. This finding gives credence to the claim made here that sub-document based approach is good for managing multi-topic documents. Since, all the algorithms developed using Java 1.6 are not highly optimized and were performed on Windows OS 64-bit platform with a 2.8 GHz CPU and 8GB memory. LDA based segmentation was less appeared than TextTiling with high computational requirements related to within document sub-document and cross sub-document clustering. For this reason, the memory costs of LDA segmentation using LDA clustering algorithm is shown in Figure 7. The results show the memory consumed for datasets (1-5). However, exact memory costs for the sub-document based clustering approach is difficult to compute. We improvised a process monitoring tool known as top for computing the memory of the LDA clustering algorithm. The results showed sub-document based clustering using the LDA segmentation and LDA algorithm is an efficient framework.

Table 9. Overview of document sets (corpus type: R-RCV1, N-20 newsgroups)

Document set	DS1	DS2	DS3	DS4	DS5
Corpus type	R	R	R -N	N	N
#documents	3290	6580	13160	26320	52640

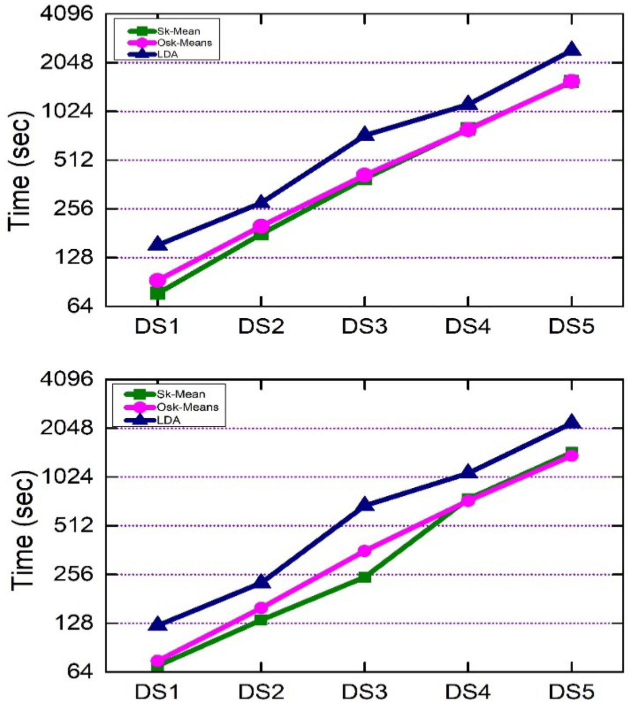


Figure 4. Time cost of within sub-document clustering using TextTiling (on top) and LDA based segmentation (on bottom)

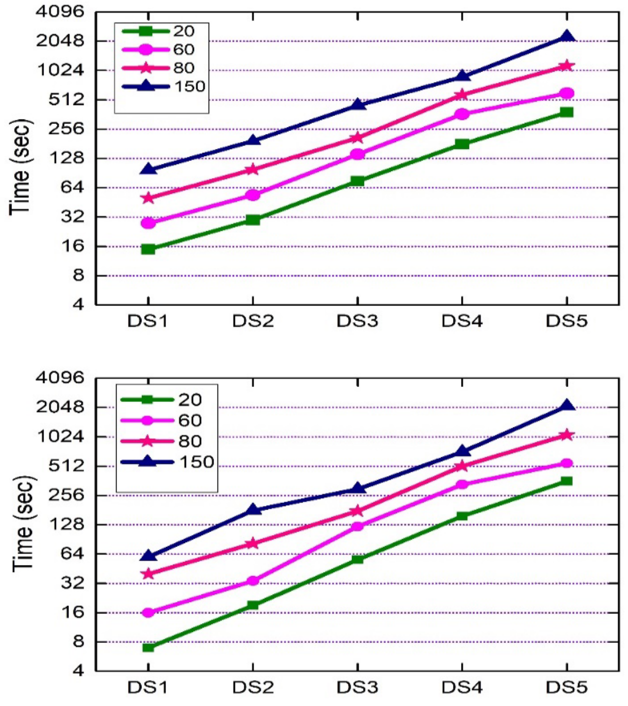


Figure 5. Time cost of cross document sub-document set clustering using TextTiling (top) and LDA based segmentation (on bottom)

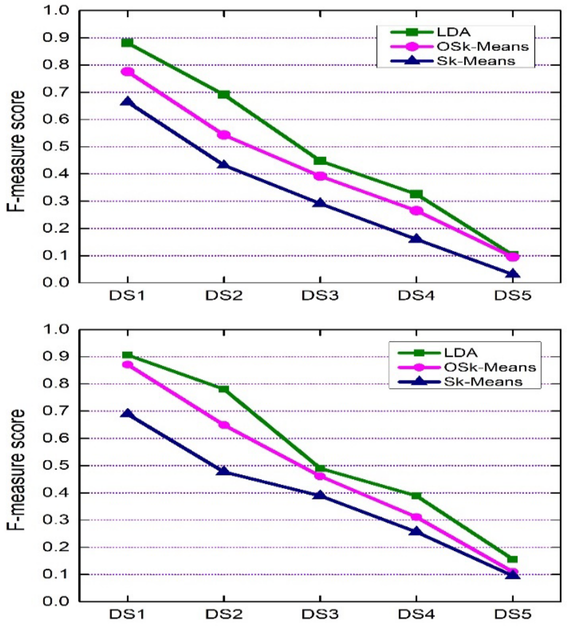


Figure 6. F-measure scores of five datasets DS (1-5) for three clustering algorithms using TextTiling (on top) and LDA based segmentation (on bottom)

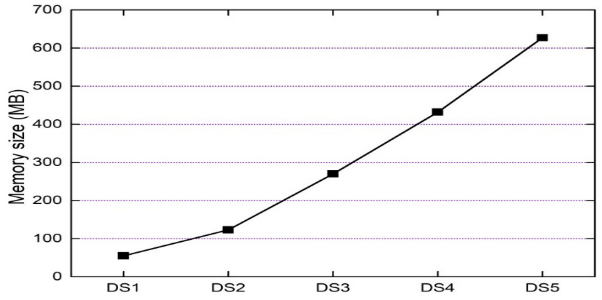


Figure 7. Memory cost of LDA based segmentation using LDA clustering algorithm for clustering documents in different datasets DS (1-5)

5.2 Experiment 2: Results and Discussions

Our proposed sub-document framework could improve performances of bisecting Sk-means and LDA. We computed algorithms on real-time data sets. Each data set was performed such as document, sub-document and sub-document set and their combination through cross and no cross clustering. Results were compared based on analysis of Sk-Means, LDA and OSk-Mean with bisecting Sk-Means and LDA. We also compared these results in TextTiling and LDA document segmentation in terms of F-measure.

5.2.1 Dataset

We derived six data set groups denoted by SET(1-6), which are extracted from 16 datasets included in four different text databases. Table 10 summarizes characteristics of all dataset groups used in experiments 2 analysis. Dataset groups denoted by SET (3-6) are extracted from Text REtrieval Conference (TREC) collections (http://trec.nist.gov) [32]. These data sets are obtained from CLUTO toolkit excluding HARD track [33]. The English newswire corpus HARD track is available in TREC containing about 650,891 documents. We stimulated same restrictions and preprocessing steps as performed and discussed in Experiment 1.

5.2.2 Document Segmentation

Document segmentation was conducted using LDA segmentation in experiment 2 and compared with the TextTiling algorithm. Table 11 and 12 shows the performance of the TextTiling and LDA, respectively. The data in Table 11 and Table 12 suggests that the higher value of the P_k and F-measure indicates more accurate segmentation.

Table 10. Datasets used in this experiment 2

Dataset	Source	#docs	#topic labels	#terms	#docs per topic	#sub-docs per doc
SET1	Classic (CACM/CISI/CRANFIELD/MEDLINE) [28]	5649	4	11219	1412.2	2.3
SET2	k1b, webkb (webACE, Web Knowledge Base) [29-30]	7440	13	33421	572.3	6.2
SET3	Ohscal (OHSUMED-233445) [31]	8639	10	10872	863.9	4.3
SET4	la1, la2, la12 (LA Times, TREC)	9181	18	86288	510.05	7.6
SET5	reviews, hitech, sports (San Jose Mercury, TREC)	11583	18	40943	643.5	5.5
SET6	HARD track (TREC)	17981	11	91921	1634.6	8.7

Table 11. Performance of TextTiling method

Data set	No cross clustering		Cross clustering		Total	
	F^μ	F^M	F^μ	F^M	F^μ	F^M
SET1	.004	.006	.013	.016	.010	.014
SET2	.005	.008	.012	.014	.009	.011
SET3	.006	.007	.010	.014	.011	.013
SET4	.004	.008	.010	.012	.008	.015
SET5	.007	.010	.011	.016	.009	.017
SET6	.006	.012	.015	.017	.012	.016

Table 12. Performance of LDA method

Data set	No cross clustering		Cross clustering		Total	
	F^μ	F^M	F^μ	F^M	p_k	F^M
SET1	.018	.025	.042	.047	18.8	0.41
SET2	.024	.030	.040	.0.43	19.3	0.40
SET3	.030	.035	.039	.044	20.3	0.38
SET4	.021	.028	.032	.046	19.1	0.36
SET5	.027	.033	.037	.044	21.5	0.37
SET6	.029	.037	.041	.043	22.2	0.39

5.2.3 Performance of Bisecting LDA using Sub-document based Framework

In the proposed sub-document based framework, clustering algorithms were performed in six different methods and presenting different sub-document representations leveraging cross and no cross clustering. In experiment 2, five clustering algorithms (Sk-mean, LDA, OSk-mean, bisecting Sk-Mean, and bisecting LDA) and their results were computed in terms of F-measure score. Further, Table 13 suggested that overlapping clustering were better than disjoint clustering. Bisecting LDA and LDA produced better performance in comparison to OSk-Means and Sk-mean via sub-document and sub-document set cross-clustering. Moreover, bisecting LDA outperforms all clustering algorithms using TextTiling method with highest macro F-measure of .739 on SET6 dataset. However, precision values on SET (1-6) were not optimal and rendered average improvements over 10%, which indicated sub-document based framework using TextTiling segmentation was slightly better than document based methods. Low recall values transpire the fact that document may relate to different domain or topics. Table 14 shows the results performed based on sub-documents segmented using LDA. Results indicated a significant improvement using bisecting LDA and Sk-Means in terms of F-measure. In addition, results suggested that bisecting LDA could performed better by producing overlapping clustering and obtained highest F-measure of 0.739 and 0.839 on SET (5-6), respectively. Bisecting Sk-mean outperformed overlapping clustering in each clustering method for all the data sets. Since, sub-document structure and topics it contain are comprised of few sentences or at least three paragraphs, hence, it may be concluded that disjoint clustering (bisecting Sk-mean) showed better performances when each document contained with few sentences in data set (SET4 and SET5). Moreover, results also suggested that overlapping clustering were far better than disjoint clustering. Table 14 transpired improved F-measure score of bisecting LDA (with an average improvement of 73%). Although, higher precision values obtained on data set SET 1 and SET (3-6), which indicated sub-document based framework using LDA segmentation assign documents properly to the clusters. However, data set SET (4,6) were obtained higher recall and lower precision values as these data sets contained huge amount of documents and involved large proportions of sub-documents with respect to the topic labels of 18 and 16 on SET (4,6) respectively.

We performed statistical significance of achieved results of proposed sub-document based framework using LDA segmentation compared to TextTiling. This significance test is assumed to use unequal variances due to multiple representation of documents (sub-document, sub-document and document) rendering higher values. Further, unpaired T test is computed through null hypothesis of no difference of achieved results i.e. entailing 60 iterations per clustering algorithm for each clustering method. Table 15 shows p values for t-test, wherein, each data set SET (1-6) shows p-values in the comparison of LDA to TextTiling for each clustering method. Also, these p-values were extremely low, which define the range of values that correspond to T-values (Two-tailed) with degree of freedom (118) at α=0.01 significance level that is equal to 2.617. These p-values indicated a clear evidence that null hypothesis is rejected. Moreover, the LDA based segmentation results were extremely significant, and therefore were superior over TextTiling in terms of F-measure.

6 Conclusion

We have presented a sub-document clustering framework integrated with LDA text segmentation. The significant contribution of this work emphasizes topic modeling to improve the segmentation methods manipulated with clustering algorithms. In addition, efficacy of the LDA text segmentation method was compared with TextTiling algorithm to detect the boundaries of the sub-documents. First, three clustering algorithms were incorporated among six clustering methods on two large datasets and compared with traditional clustering using hard and soft partitional clustering algorithms. Later, bisecting clustering algorithms were performed on real time data sets derived from multiple sources. The results obtained by our proposed framework suggested that our approach outperformed the clustering methods such as segment-based framework and traditional clustering methods for multi-document, is hence significant in terms of F-measure particularly using LDA bisecting clustering algorithm and LDA segmentation method. To the best of our knowledge, sub-document based clustering using the LDA based segmentation framework considerably improved the recognition of different topics within a document. The proposed framework can be helpful in document segment detection and to control their length when documents are paragraph-less and to produce algorithms for document identification where the topical structure of documents is required. In addition, it can also be incorporated in topic detection and novelty detection to distinguish and recognize topically coherent sub-documents in a document. We further extend our work by incorporating ensemble and ontology based clustering methods embedding to our sub-document based framework for sparse and high dimensional data in order to reduce the computational complexity.

Table 13. Results performance of sub-document based clustering using TextTiling

Clustering Method	Clustering Algorithm	Dataset																	
		SET1			SET2			SET3			SET4			SET5			SET6		
		P	R	F^M	P	R	F^M	P	R	F^M	P	R	F^M	P	R	F^M	P	R	F^M
No cross document clustering	Sk-Means	0.592	0.157	0.248	0.45	0.24	0.31	0.648	0.13	0.217	0.282	0.767	0.412	0.617	0.527	0.568	0.55	0.371	0.443
	LDA	0.169	0.64	0.267	0.25	0.62	0.36	0.371	0.51	0.43	0.317	0.604	0.416	0.57	0.482	0.522	0.47	0.56	0.513
	OSk-Means	0.142	0.534	0.224	0.22	0.43	0.29	0.314	0.422	0.36	0.337	0.69	0.453	0.583	0.362	0.447	0.44	0.541	0.483
	Bisecting Sk-Means	0.563	0.508	0.534	0.6	0.28	0.38	0.733	0.124	0.212	0.396	0.814	0.533	0.695	0.551	0.615	0.64	0.404	0.495
	Bisecting LDA	0.532	0.61	0.568	0.35	0.59	0.44	0.478	0.56	0.516	0.394	0.691	0.502	0.632	0.434	0.515	0.61	0.621	0.616
No cross sub-document clustering	Sk-Means	0.612	0.217	0.32	0.47	0.3	0.36	0.668	0.19	0.296	0.302	0.827	0.442	0.637	0.587	0.611	0.57	0.431	0.491
	LDA	0.209	0.671	0.319	0.29	0.65	0.4	0.411	0.541	0.467	0.357	0.635	0.457	0.61	0.513	0.557	0.51	0.591	0.55
	OSk-Means	0.152	0.536	0.237	0.23	0.43	0.3	0.324	0.424	0.367	0.347	0.692	0.462	0.593	0.364	0.451	0.45	0.543	0.49
	Bisecting Sk-Means	0.593	0.518	0.553	0.63	0.29	0.4	0.763	0.134	0.228	0.426	0.824	0.562	0.725	0.561	0.633	0.67	0.414	0.511
	Bisecting LDA	0.552	0.645	0.595	0.37	0.62	0.46	0.498	0.595	0.542	0.414	0.726	0.527	0.652	0.469	0.546	0.63	0.656	0.643
Sub-document cross clustering	Sk-Means	0.643	0.227	0.336	0.5	0.31	0.38	0.699	0.2	0.311	0.333	0.837	0.476	0.668	0.597	0.631	0.6	0.441	0.509
	LDA	0.253	0.71	0.373	0.34	0.69	0.45	0.455	0.58	0.51	0.401	0.674	0.503	0.654	0.552	0.599	0.56	0.63	0.592
	OSk-Means	0.229	0.625	0.335	0.31	0.52	0.39	0.401	0.513	0.45	0.424	0.781	0.55	0.67	0.453	0.541	0.52	0.632	0.572
	Bisecting Sk-Means	0.573	0.588	0.58	0.61	0.36	0.45	0.743	0.204	0.32	0.406	0.894	0.558	0.705	0.631	0.666	0.65	0.484	0.554
	Bisecting LDA	0.583	0.73	0.648	0.4	0.71	0.51	0.529	0.68	0.595	0.445	0.811	0.575	0.683	0.554	0.612	0.66	0.741	0.699
Document cross clusterin	Sk-Means	0.631	0.216	0.322	0.49	0.3	0.37	0.687	0.189	0.296	0.321	0.826	0.462	0.656	0.492	0.562	0.59	0.492	0.537
	LDA	0.241	0.689	0.357	0.32	0.67	0.44	0.443	0.559	0.494	0.389	0.653	0.488	0.642	0.531	0.581	0.55	0.609	0.576
	OSk-Means	0.199	0.605	0.299	0.28	0.5	0.36	0.371	0.493	0.423	0.394	0.761	0.519	0.64	0.433	0.517	0.49	0.612	0.546
	Bisecting Sk-Means	0.557	0.538	0.547	0.59	0.31	0.41	0.727	0.154	0.254	0.39	0.844	0.533	0.689	0.581	0.63	0.63	0.434	0.515
	Bisecting LDA	0.532	0.709	0.608	0.35	0.69	0.46	0.478	0.659	0.554	0.394	0.79	0.526	0.632	0.533	0.578	0.61	0.72	0.661
Sub-document set cross clustering	Sk-Means	0.654	0.257	0.369	0.51	0.34	0.41	0.71	0.23	0.347	0.344	0.867	0.493	0.679	0.627	0.652	0.61	0.471	0.533
	LDA	0.317	0.74	0.444	0.4	0.72	0.51	0.519	0.61	0.561	0.465	0.704	0.56	0.718	0.582	0.643	0.62	0.66	0.64
	OSk-Means	0.239	0.635	0.347	0.32	0.53	0.4	0.411	0.523	0.46	0.434	0.791	0.56	0.68	0.463	0.551	0.53	0.642	0.582
	Bisecting Sk-Means	0.593	0.558	0.575	0.63	0.33	0.43	0.763	0.174	0.283	0.426	0.864	0.571	0.725	0.601	0.657	0.67	0.454	0.541
	Bisecting LDA	0.603	0.74	0.665	0.42	0.72	0.53	0.549	0.69	0.611	0.465	0.821	0.594	0.703	0.564	0.626	0.68	0.751	0.715
Sub-document & sub-document set cross clustering	Sk-Means	0.675	0.267	0.383	0.53	0.35	0.42	0.731	0.24	0.361	0.365	0.877	0.515	0.7	0.637	0.667	0.63	0.481	0.547
	LDA	0.349	0.76	0.478	0.43	0.74	0.55	0.551	0.63	0.588	0.4965	0.724	0.589	0.75	0.602	0.668	0.65	0.68	0.666
	OSk-Means	0.26	0.665	0.374	0.34	0.56	0.42	0.432	0.553	0.485	0.455	0.821	0.586	0.701	0.493	0.579	0.55	0.672	0.607
	Bisecting Sk-Means	0.663	0.538	0.594	0.7	0.31	0.43	0.833	0.154	0.26	0.496	0.844	0.625	0.775	0.581	0.664	0.72	0.434	0.541
	Bisecting LDA	0.615	0.78	0.688	0.43	0.76	0.55	0.561	0.73	0.634	0.477	0.861	0.614	0.715	0.604	0.655	0.69	0.791	0.739

Table 14. Results performances of sub-document based clustering using LDA segmentation

Clustering Method	Clustering Algorithm	Dataset																	
		SET1			SET2			SET3			SET4			SET5			SET6		
		P	R	F^M	P	R	F^M	P	R	F^M	P	R	F^M	P	R	F^M	P	R	F^M
No cross document clustering	Sk-Means	0.674	0.24	0.354	0.533	0.25	0.34	0.699	0.16	0.26	0.829	0.306	0.447	0.492	0.594	0.538	0.603	0.433	0.504
	LDA	0.235	0.716	0.354	0.33	0.67	0.442	0.486	0.576	0.527	0.347	0.584	0.435	0.609	0.683	0.644	0.574	0.64	0.605
	OSk-Means	0.21	0.576	0.308	0.274	0.479	0.349	0.361	0.473	0.409	0.349	0.711	0.468	0.597	0.516	0.554	0.472	0.596	0.527
	Bisecting Sk-Means	0.58	0.529	0.553	0.59	0.328	0.422	0.801	0.227	0.354	0.883	0.333	0.484	0.592	0.592	0.592	0.665	0.431	0.523
	Bisecting LDA	0.567	0.729	0.638	0.447	0.662	0.534	0.636	0.585	0.609	0.485	0.7	0.573	0.652	0.681	0.666	0.692	0.745	0.718
No cross sub-document clustering	Sk-Means	0.694	0.3	0.419	0.553	0.31	0.397	0.719	0.22	0.337	0.849	0.366	0.511	0.512	0.654	0.574	0.623	0.493	0.55
	LDA	0.275	0.747	0.402	0.37	0.701	0.484	0.526	0.607	0.564	0.387	0.615	0.475	0.649	0.714	0.68	0.614	0.671	0.641
	OSk-Means	0.22	0.578	0.319	0.284	0.481	0.357	0.371	0.475	0.417	0.359	0.713	0.478	0.607	0.518	0.559	0.482	0.598	0.534
	Bisecting Sk-Means	0.61	0.539	0.572	0.62	0.338	0.437	0.831	0.237	0.369	0.913	0.343	0.499	0.622	0.602	0.612	0.695	0.441	0.54
	Bisecting LDA	0.587	0.764	0.664	0.467	0.697	0.559	0.656	0.62	0.637	0.505	0.735	0.599	0.672	0.716	0.693	0.712	0.78	0.744
Sub-document cross clustering	Sk-Means	0.725	0.31	0.434	0.584	0.32	0.413	0.75	0.23	0.352	0.88	0.376	0.527	0.543	0.664	0.597	0.654	0.503	0.569
	LDA	0.319	0.786	0.454	0.414	0.74	0.531	0.57	0.646	0.606	0.431	0.654	0.52	0.693	0.753	0.722	0.658	0.71	0.683
	OSk-Means	0.297	0.667	0.411	0.361	0.57	0.442	0.448	0.564	0.499	0.436	0.802	0.565	0.684	0.607	0.643	0.559	0.687	0.616
	Bisecting Sk-Means	0.59	0.609	0.599	0.6	0.408	0.486	0.811	0.307	0.445	0.893	0.413	0.565	0.602	0.672	0.635	0.675	0.511	0.582
	Bisecting LDA	0.618	0.849	0.715	0.498	0.782	0.608	0.687	0.705	0.696	0.536	0.82	0.648	0.703	0.801	0.749	0.743	0.865	0.799
Document cross clustering	Sk-Means	0.713	0.299	0.421	0.572	0.309	0.401	0.738	0.219	0.338	0.868	0.492	0.628	0.531	0.492	0.511	0.642	0.492	0.557
	LDA	0.307	0.765	0.438	0.402	0.719	0.516	0.558	0.625	0.59	0.419	0.633	0.504	0.681	0.732	0.706	0.646	0.689	0.667
	OSk-Means	0.267	0.647	0.378	0.331	0.55	0.413	0.418	0.544	0.473	0.406	0.782	0.534	0.654	0.587	0.619	0.529	0.667	0.59
	Bisecting Sk-Means	0.574	0.559	0.566	0.584	0.358	0.444	0.795	0.257	0.388	0.877	0.363	0.513	0.586	0.622	0.603	0.659	0.461	0.542
	Bisecting LDA	0.567	0.828	0.673	0.447	0.761	0.563	0.636	0.684	0.659	0.485	0.799	0.604	0.652	0.78	0.71	0.692	0.844	0.76
Sub-document set cross clustering	Sk-Means	0.736	0.34	0.465	0.5952	0.35	0.441	0.7612	0.26	0.388	0.8912	0.406	0.558	0.5542	0.694	0.616	0.6652	0.533	0.592
	LDA	0.383	0.816	0.521	0.478	0.77	0.59	0.634	0.676	0.654	0.495	0.684	0.574	0.757	0.783	0.77	0.722	0.74	0.731
	OSk-Means	0.307	0.677	0.422	0.371	0.58	0.453	0.458	0.574	0.509	0.446	0.812	0.576	0.694	0.617	0.653	0.569	0.697	0.627
	Bisecting Sk-Means	0.61	0.579	0.594	0.62	0.378	0.47	0.831	0.277	0.416	0.913	0.383	0.54	0.622	0.642	0.632	0.695	0.481	0.569
	Bisecting LDA	0.638	0.859	0.732	0.518	0.792	0.626	0.707	0.715	0.711	0.556	0.83	0.666	0.723	0.811	0.764	0.763	0.875	0.815
Sub-document & sub-document set cross clustering	Sk-Means	0.757	0.35	0.479	0.6162	0.36	0.454	0.7822	0.27	0.401	0.9122	0.416	0.571	0.5752	0.704	0.633	0.6862	0.543	0.606
	LDA	0.415	0.836	0.555	0.5095	0.79	0.619	0.6655	0.696	0.68	0.5265	0.704	0.602	0.7885	0.803	0.796	0.7535	0.76	0.757
	OSk-Means	0.328	0.707	0.448	0.392	0.61	0.477	0.479	0.604	0.534	0.467	0.842	0.601	0.715	0.647	0.679	0.59	0.727	0.651
	Bisecting Sk-Means	0.68	0.559	0.614	0.69	0.358	0.471	0.901	0.257	0.4	0.963	0.363	0.527	0.672	0.622	0.646	0.745	0.461	0.57
	Bisecting LDA	0.65	0.899	0.754	0.53	0.832	0.648	0.719	0.755	0.737	0.568	0.87	0.687	0.735	0.851	0.789	0.775	0.915	0.839

Table 15. P-values for unpaired T-test (df=118)

Data set	LDA segmentation method versus TextTiling					
	No Cross Document Clustering	No Cross Sub-document Clustering	Sub-document Cross Clustering	Document Cross Clustering	Sub-document Set Cross Clustering	Sub-document and sub-document set Cross Clustering
SET1	2.16E-38	4.46E-48	3.33E-22	5.13E-16	3.89E-13	5.22E-18
SET2	4.11E-42	5.66E-33	2.55E-18	7.22E-12	2.11E-10	4.25E-15
SET3	3.33E-51	5.10E-60	5.57E-26	3.92E-23	9.11E-19	8.10E-24
SET4	1.21E-21	6.12E-25	4.32E-8	5.21E-11	1.73E-17	2.11E-14
SET5	1.44E-12	2.91E-16	3.11E-27	2.18E-20	7.19E-18	5.25E-17
SET6	1.13E-69	4.10E-48	3.16E-34	4.36E-28	5.15E-29	2.59E-31

Acknowledgements

The work in this paper has been supported by National Natural Science Foundation of China (61602456), National High-tech R&D Program (863 Program) (2015AA017204).

References

[1] A. Bouguettaya, Qi. Yu, X. Liu, X. Zhou, A. Song, Efficient Agglomerative Hierarchical Clustering, *Expert Systems with Applications*, Vol. 42, No. 5, pp. 2785-2797, April, 2015.

[2] A. Castellanos, J. Cigarrán, A. García-Serrano, Formal Concept Analysis for Topic Detection: A Clustering Quality Experimental Analysis, *Information Systems*, Vol. 66, pp. 24-42, June, 2017.

[3] H. Zhao, S. Salloum, Y. Cai, J. Z. Huang, Ensemble Subspace Clustering of Text Data Using Two-Level Features, *International Journal of Machine Learning and Cybernetics*, Vol. 8, No. 6, pp. 1-16, December, 2017.

[4] T.-C. Chang, H. Wang, S. Yu, A Novel Approach for Complex Datasets Clustering/Classification, *Journal of Internet Technology*, Vol. 17, No. 3, pp. 523-530, May, 2016.

[5] K. Niu, Z. Gao, H. Jiao, X. Qiao, Y. Zhao, Subspace Clustering for Vector Clusters, *Journal of Internet Technology*, Vol. 18, No. 1, pp. 87-94, January, 2017.

[6] A. Tagarelli, G. Karypis, Segment-Based Approach to Clustering Multi-Topic Documents, *Proceedings of the Sixth Workshop on Text Mining, in Conjunction with the 8th SIAM International Conference on Data Mining (SDM)*, USA, January, 2008, pp. 1-12.

[7] Tagarelli, G. Karypis. A Segment-Based Approach to Clustering Multi-Topic Documents, *Knowledge and Information Systems*, Vol. 34, No. 3, pp. 563-595, March, 2013.

[8] A. Kelaiaia, H. Merouani, Clustering with Probabilistic Topic Models on Arabic Texts: A Comparative Study of LDA and K-Means, *International Arab Journal of Information Technology*, Vol. 13, No. 2, pp. 332-338, March, 2016.

[9] T. Hofmann, Unsupervised Learning by Probabilistic Latent Semantic Analysis, *Machine Learning*, Vol. 42, No. 1-2, pp. 177-196, January, 2001.

[10] T. Hofmann, Probabilistic Latent Semantic Indexing. *Proceedings of the 22nd ACM International Conference on Research and Development in Information Retrieval (SIGIR'99)*, Berkeley, CA, 1999, pp. 50-57.

[11] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet Allocation, *Journal of Machine Learning*, Vol. 3, pp. 993-1022, January, 2003.

[12] F. Y. Y. Choi, Advances in Domain Independent Linear Text Segmentation, *Proceedings of the Conference of 1st North American Chapter of the Association for Computational Linguistics Conference (NAACL 2000)*, Seattle, WA, 2000, pp. 26-33.

[13] C.-K. Yau, A. Porter, N. Newman, A. Suominen, Clustering Scientific Documents with Topic Modeling. *Scientometrics*, Vol. 100, No. 3, pp. 767-786, September, 2014.

[14] S. Nicola, C. Joe, S. F Alan, Select a Lexical Cohesion Based News Story Segmentation System, *AI Communications*, Vol. 17, No. 1, pp. 3-12, January, 2004.

[15] Y. Bestgen, Improving Text Segmentation Using Latent Semantic Analysis: A Reanalysis of Choi, Wiemer-Hastings, and Moore, *Computational Linguistics*, Vol. 32, No. 1, pp. 5-12, May, 2006.

[16] M. Ana, C. Andrea, An Ontology-based Approach to Information Retrieval, in: A. Cali, D. Gorgan, M. Ugarte (Eds.), *Semantic Keyword-Based Search on Structured Data Sources*, Lecture Notes in Computer Science, Vol. 10151, Springer, 2017, pp. 150-156.

[17] A. Sara, T. Fattaneh, A Semantic ontology-based Document Organizer to Cluster Elearning Documents, *Second International Conference on Web Research (ICWR)*, Tehran, Iran, 2017, pp. 1-7.

[18] T. Wei, Y. Lu, H. Chang, Q. Zhou, X. Bao, A Semantic Approach for Text Clustering Using WordNet and Lexical Chains, *Expert Systems with Applications*, Vol. 42, No. 4, pp. 2264-2275, March, 2015.

[19] S. Niusha, L. H. Lee, R. Rajkumar, V. P. Kallimani, A. N. Ahmed, Using Unsupervised Clustering Approach to Train the Support Vector Machine for Text Classification, *Neurocomputing*, Vol. 211, pp. 4-10, October, 2016.

[20] T. Volkan, b. Turgay, C. Ali, An Improved Text Clustering Algorithm for Text Mining: Multi-Cluster Spherical K-Means, *International Arab Journal of Information Technology*, Vol. 13, No. 1, pp. 12-19, January, 2016.

[21] Z. Mohammed, D. A. Baraani, A. Ehsan, S. Alireza, K. A. Akhavan, A New Experience in Persian Text Clustering using FarsNet Ontology, *Journal of Information Science and*

Engineering, Vol. 31, no. 1, pp. 315-330, January, 2015.

[22] A. Rifki, K. Retno, G. Rahmat, Topic Labelling Towards News Document Collection Based on Latent Dirichlet Allocation and Ontology, *Proceeding of the first International Conference on Informatics and Computational Sciences (ICICOS)*, Semarang, Indonesia, 2017, pp. 247-251.

[23] L. Yaxiong, P. Deng, Text Clustering Based on Domain Ontology and Latent Semantic Analysis, *Proceeding of the International Conference of the Mechatronics Engineering and Computing Technology (ICMECT)*, Shanghai, China, 2014, pp. 556-562.

[24] L. Yue, W. Zuo, T. Peng, Y. Wang, X. Han, A Fuzzy Document Clustering Approach Based on Domain-Specified Ontology, *Data and Knowledge Engineering*, Vol. 100, pp. 148-166, November, 2015.

[25] J. Ahmed, F. Mohamed, F. Mohamed, Enhanced Clustering-based Topic Identification of Transcribed Arabic Broadcast News, *International Arab Journal of Information Technology*, Vol. 14, No. 15, pp. 721-728, September, 2017.

[26] D. D. Lewis, Y. Yang, T. G. Rose, F. Li, RCV1: A New Benchmark Collection for Text Categorization Research, *Journal of Machine Learning Research*, Vol. 5, pp. 361-397, December, 2004.

[27] Rennie, The 20 Newsgroups Data Set, http://qwone.com/jason/20 Newsgroups/.

[28] Classic Text Database, ftp://ftp.cs.cornell.edu/pub/smart/.

[29] E. H. Han, D. Boley, M. Gini, R. Gross, K. Hastings, G. Karypis, V. Kumar, B. Mobasher, J. Moore, WebACE: A Web Agent for Document Categorization and Exploration, *Proceedings of the 2nd International Conference Autonomous Agents*, 1998, pp. 408-415.

[30] M. Craven, D. DiPasquo, A. McCallum, T. Mitchell, K. Nigam, S. Slattery, Learning to Extract Symbolic Knowledge from the World Wide Web, *the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence (AAAI-98/IAAI-98)*, Madison, WI, 1998, pp. 509-516.

[31] Hersh, C. Buckley, T. J. Leone, D. Hickam, OHSUMED: An Interactive Retrieval Evaluation and New Large Test Collection for Research, *the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Dublin, Ireland, 1994, pp. 192-201.

[32] E. Voorhees, D. Harman, The Text Retrieval Conferences (TRECS*), Proceedings of a Workshop*, Baltimore, Maryland, 1998, pp. 241-273.

[33] Y. Zhao, G. Karypis, Criterion Functions for Document Clustering: Experiments and Analysis, Technical Report #01-40, 2001.

Biographies

Muhammad Qasim Memon is currently a post-doctorate fellow in Advanced Innovation Center for Future Education (AICFE), Faculty of Education, Beijing Normal University in China. He received his B.E. and M.E. degrees from Mehran University of Engineering & Technology Jamshoro (MUET). He received his Ph.D. degree at school of Software Engineering from Beijing University of Technology in China. Dr. Memon's research interests include text mining, information extraction and wireless sensor networks.

Jingsha He is currently a professor in the Faculty of Information Technology, Beijing University of Technology in China. He received his B.S. degree from Xi'an Jiaotong University in China and his M.S. and Ph.D. degrees from University of Maryland at College Park in U.S. Prof. He's research interests include information security, wireless networks and digital forensics.

Yu Lu received his Ph.D. degree from the National University of Singapore. He is currently an Associate Professor with the Faculty of Education, Beijing Normal University, where he also serves as the director of the artificial intelligence (AI) lab and leads the research team for AI in education. His recent research interests include educational data mining, learning analytics, pervasive computing and educational robotics.

Nafei Zhu is currently a lecturer in the Faculty of Information Technology, Beijing University of Technology in China. She received her B.S. and M.S. degrees from Central South University in China and her Ph.D. degree from Beijing University of Technology in China. Ms. Zhu's research interests include information security, privacy and network measurement.

Aasma Memon received her B.A. and M.P.A. degrees from University of Sindh, Jamshoro, Pakistan in 2008 and 2012 respectively. She is currently a Ph.D. scholar at school of Economics and Management in Beijing University of Technology. Her research interests include management information system, human resource management and data mining.